BrainState-ML: Deciphering Cognitive Assessment Performance and Brain Functional Organization from Task-fMRI Using Interpretable Machine Learning

Eason Zhou
Glenview Senior Public School, Canada

Dr. Haoyan Jiang
FutureEra Research Institute, Canada

Project Track: Natural & Applied Sciences - Bio-informatics

Country / Region: Canada

June 26, 2026

***Abstract*—Task-based functional magnetic resonance imaging (fMRI) provides a means of examining how cognitive states are represented through distributed brain activation patterns. This paper presents an interpretable machine learning framework for decoding cognitive task states from task-fMRI contrast maps using the Brainomics Localizer dataset. The analysis includes 30 participants and four cognitive task categories: Working Memory / Math, Language / Reading, Motor Control, and Visual Processing. Each three-dimensional contrast map is transformed into a 100-dimensional regional activation vector using the Schaefer 2018 cortical atlas, enabling compact and interpretable representation of task-evoked cortical activity. Four supervised classifiers—logistic regression, k-nearest neighbors, decision tree, and random forest—are optimized through stratified cross-validation and evaluated using accuracy, precision, recall, F1-score, and multiclass AUROC. Experimental results show that task states can be decoded reliably from regional activation features. Random Forest achieves the highest classification accuracy of 85.83%, while Logistic Regression achieves comparable accuracy of 85.00% and the highest macro AUROC of 0.9656. Because of its strong probabilistic separability and direct coefficient interpretability, Logistic Regression is selected for network-level analysis. The confusion matrix shows that Motor Control and Visual Processing are the most separable categories, whereas most errors occur between Working Memory / Math and Language / Reading. Feature-importance analysis further maps Schaefer regional coefficients onto Yeo seven-network systems, revealing task-relevant contributions from control, attention, somatomotor, and visual networks. Overall, the proposed framework demonstrates that atlas-based machine learning can support both cognitive task decoding and interpretable functional brain mapping from task-fMRI contrast data.**

I. Introduction

Task-based functional magnetic resonance imaging (fMRI) provides a powerful method for examining how cognitive operations are represented in the human brain. During controlled experimental tasks, fMRI measures blood-oxygen-level-dependent responses that reflect task-evoked neural activity. By comparing activation patterns across task conditions, researchers can investigate how different cognitive and sensorimotor processes are distributed across cortical systems [1]. Traditional task-fMRI analysis commonly relies on general linear models, statistical contrast maps, and group-level activation inference. These methods remain central to hypothesis-driven neuroimaging, but they are less directly suited to determining whether distributed activation patterns can predict the cognitive state from which they were generated [2].

Machine learning offers a complementary approach by treating task-fMRI analysis as a brain-decoding problem. Rather than asking only which regions are significantly activated, decoding models evaluate whether multivariate brain activation profiles contain sufficient information to classify task conditions [3], [4], [5]. This perspective is particularly relevant for cognitive neuroscience because cognitive functions are not localized to single isolated regions; they emerge from distributed interactions among sensory, motor, attentional, executive-control, and association networks. However, predictive separability must be interpreted cautiously, since successful classification does not by itself establish a direct psychological or causal explanation of the decoded cognitive construct [6].

In this study, we develop an interpretable machine learning framework for decoding cognitive task states from task-fMRI contrast maps. The analysis uses the Brainomics Localizer dataset accessed through the `nilearn` neuroimaging library [5], [7]. The implemented dataset contains statistical activation maps from 30 participants across four selected task contrasts, producing 120 total contrast maps. The decoded task categories are *Working Memory / Math*, *Language / Reading*, *Motor Control*, and *Visual Processing*, corresponding respectively to visual calculation, sentence reading, left-versus-right button pressing, and horizontal-versus-vertical checkerboard stimulation.

To obtain compact and interpretable neuroimaging features, each three-dimensional contrast map is transformed into a 100-dimensional regional activation vector using the Schaefer 2018 cortical atlas [8]. Each feature represents the mean activation value within one Schaefer cortical parcel. These parcels are further associated with the Yeo seven-network organization, which groups cortical regions into canonical large-scale functional systems, including visual, somatomotor, dorsal-attention, ventral-attention, limbic, frontoparietal-control, and default-mode networks [9]. This representation reduces voxel-level dimensionality while preserving a direct link between machine learning features and interpretable brain systems.

The resulting problem is formulated as a balanced four-class cognitive decoding task. The study addresses three questions: first, whether Schaefer-100 regional activation features can distinguish among cognitive task states; second, how different supervised classifiers compare under stratified cross-validation; and third, which cortical networks contribute most strongly to task discrimination. To answer these questions, we benchmark logistic regression, k-nearest neighbors, decision tree, and random forest classifiers using accuracy, precision, recall, F1-score, and multi-class AUROC [10]. We further interpret the selected model by aggregating regional feature-importance values into Yeo seven-network profiles, thereby connecting predictive task decoding with large-scale functional brain organization.

The main contribution of this work is an end-to-end, interpretable pipeline for task-fMRI cognitive decoding. By combining Brainomics Localizer contrast maps, Schaefer-100 regional feature extraction, comparative supervised learning, and Yeo-network interpretation, the proposed framework provides a practical method for classifying cognitive task states while retaining neuroscientific interpretability.

II. Methods

A. Dataset and Cognitive Task Contrasts

This study used the Brainomics Localizer functional neuroimaging dataset accessed through the `nilearn.datasets.fetch_localizer_contrasts` interface [11], [12]. Rather than using raw task-fMRI time series, the present implementation operates on precomputed statistical contrast maps, where each image summarizes task-evoked activation associated with a specific cognitive condition. This design makes the pipeline computationally lightweight while preserving spatially distributed neural signatures that are informative for cognitive-state decoding.

The implemented analysis included 30 participants and four task-related contrasts, yielding a balanced dataset of 120 activation maps. Each participant contributed one statistical map for each contrast. The four contrasts were selected to represent distinct cognitive and sensorimotor domains: visual calculation, sentence reading, lateralized button pressing, and checkerboard orientation processing. These contrasts were subsequently relabeled into four interpretable cognitive categories: *Working Memory / Math*, *Language / Reading*, *Motor Control*, and *Visual Processing*. The visual calculation contrast was treated as a mathematical cognition condition involving symbolic manipulation and executive processing; the sentence-reading contrast represented language comprehension; the left-versus-right button-press contrast represented motor control; and the horizontal-versus-vertical checkerboard contrast represented low-level visual processing.

Formally, the dataset can be expressed as

$$\mathcal{D} = \{(\mathbf{I}_i, y_i)\}_{i=1}^{N},$$

where $\mathbf{I}_i$ denotes a three-dimensional statistical activation map, y_i denotes its cognitive task label, and $N = 120$. Since the four classes were equally represented, each category contained 30 observations. The resulting classification problem was therefore a balanced four-class cognitive decoding task.

B. Regional Activation Feature Extraction

To convert each three-dimensional activation map into a compact machine learning representation, the brain was parcellated using the Schaefer 2018 cortical atlas [8], [9] with 100 regions of interest and 7 canonical Yeo functional networks. The atlas was fetched at 2 mm spatial resolution, and each region label was parsed to recover its associated Yeo network assignment, including visual, somatomotor, dorsal-attention, ventral-attention, limbic, frontoparietal-control, and default-mode systems.

For each contrast map, the `NiftiLabelsMasker` was used to compute the mean activation value inside each Schaefer parcel. This operation transformed each whole-brain map into a 100-dimensional regional activation vector:

$$\mathbf{x}_i = [a_{i1}, a_{i2}, \dots, a_{i100}],$$

where a_{ij} denotes the average activation value of image i within region j. The complete feature matrix therefore had shape

$$\mathbf{X} \in \mathbb{R}^{120\times 100},$$

with rows corresponding to subject–contrast activation maps and columns corresponding to Schaefer atlas regions. The target vector had shape

$$\mathbf{y} \in \{1, 2, 3, 4\}^{120},$$

corresponding to the four decoded cognitive categories.

This feature construction intentionally emphasizes regional interpretability. Instead of training directly on voxel-level images, which would produce a very high-dimensional feature space, the atlas-based representation compresses each activation map into anatomically organized cortical summaries. This reduces the risk of overfitting, improves computational efficiency, and permits subsequent interpretation at both regional and functional-network levels.

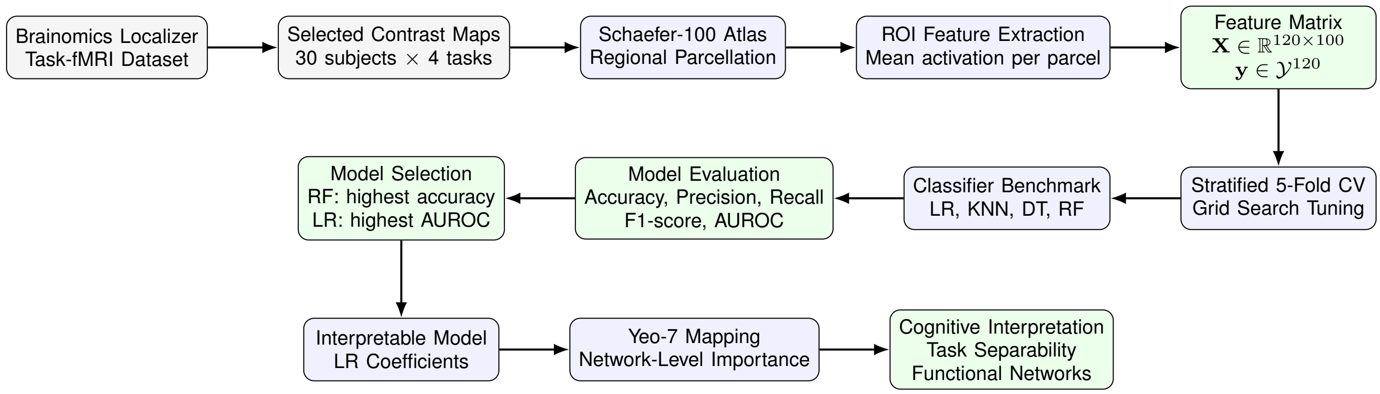

Figure 1. Project pipeline for interpretable cognitive task decoding from task-fMRI contrast maps. Brainomics Localizer activation maps are converted into Schaefer-100 regional features, evaluated through stratified cross-validation across four classifiers, and interpreted using logistic-regression coefficients aggregated into Yeo seven-network functional systems.

C. Cognitive-State Prediction Task

The primary prediction objective was multiclass cognitive-state decoding. Given a 100-dimensional vector of regional brain activations, the model was trained to predict which cognitive task category generated the corresponding activation map. The task can be formulated as

$$\mathcal{Y} = \left\{ \begin{array}{l} \text{Working Memory / Math,} \\ \text{Language / Reading,} \\ \text{Motor Control,} \\ \text{Visual Processing} \end{array} \right\}.$$

The decoding objective is therefore formulated as

$$f : \mathbb{R}^{100} \to \mathcal{Y}.$$

This formulation differs from behavioral cognitive-performance prediction. The current implementation does not predict reaction time, accuracy, error rate, or demographic variables. Instead, it asks whether distributed cortical activation patterns contain sufficient information to distinguish among cognitive task domains. The analysis therefore focuses on decoding the functional brain state associated with each task condition and subsequently interpreting which brain networks support the classifier's decisions.

"'latex

D. Classifier Architectures and Hyperparameter Optimization

The final classification framework was designed as a comparative benchmark across four supervised machine learning classifiers: multinomial logistic regression, k-nearest neighbors, decision tree, and random forest. These models were selected to represent complementary inductive assumptions. Logistic regression provides a linear and interpretable baseline; k-nearest neighbors evaluates whether task categories are separable through local similarity in the regional activation space; decision trees provide a nonlinear rule-based classifier; and random forests extend decision trees through ensemble aggregation to improve robustness and reduce variance.

Given the regional activation matrix

$$\mathbf{X} \in \mathbb{R}^{120\times 100}$$

and the task-label vector

$$\mathbf{y} \in \mathcal{Y}^{120},$$

each classifier was trained to approximate the mapping

$$f : \mathbb{R}^{100} \to \mathcal{Y},$$

where $\mathcal{Y}$ denotes the four cognitive task categories.

For each model family, hyperparameters were optimized using stratified k-fold cross-validation. Stratification ensured that each fold preserved the balanced distribution of the four cognitive task classes. The optimization process searched over model-specific hyperparameter grids and selected the configuration that maximized cross-validated classification performance on the training folds.

For multinomial logistic regression, the principal tuned hyperparameter was the inverse regularization strength C, with L2 regularization used to control coefficient magnitude and reduce overfitting. For k-nearest neighbors, the number of neighbors, distance metric, and neighbor-weighting scheme were tuned. For the decision tree classifier, the maximum tree depth, split criterion, minimum number of samples required for a split, and minimum number of samples per leaf were optimized. For the random forest classifier, the number of trees, maximum depth, number of candidate features per split, and leaf-size constraints were tuned.

The candidate model families can be summarized as

$$\mathcal{M} = \left\{ \begin{array}{l} \text{Logistic Regression,} \\ \text{k-Nearest Neighbors,} \\ \text{Decision Tree,} \\ \text{Random Forest} \end{array} \right\}.$$

For each model $m \in \mathcal{M}$, the best hyperparameter setting was selected as

$$\theta_m^* = \arg\max_{\theta_m} \mathrm{CVScore}(m_{\theta_m}, \mathbf{X}, \mathbf{y}),$$

where θ_m denotes the hyperparameter configuration for model m, and $\mathrm{CVScore}$ denotes the stratified cross-validation score used during model selection.

Because logistic regression and k-nearest neighbors are sensitive to feature scale, regional activation features were standardized inside the cross-validation pipeline. This ensured that scaling parameters were estimated only from training folds and then applied to the corresponding validation folds, thereby avoiding information leakage. Tree-based models are scale-invariant, but they were evaluated under the same cross-validation protocol to ensure a consistent benchmarking procedure.

E. Validation Strategy and Benchmark Metrics

After hyperparameter tuning, the optimized classifiers were benchmarked using stratified cross-validation. The primary evaluation objective was to compare the ability of each model to decode cognitive task categories from Schaefer-100 regional activation features [13]. Since the dataset contained an equal number of samples for each of the four task categories, the benchmark emphasized both overall classification performance and class-specific behavior.

Model performance was evaluated using accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve. Accuracy measured the overall proportion of correctly classified activation maps:

$$\mathrm{Accuracy} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(\hat{y}_i = y_i).$$

Precision, recall, and F1-score were computed for each cognitive task class and then summarized using macro-averaging. Macro-averaging was used because it gives equal weight to each task category, making it suitable for evaluating whether the classifier performs consistently across working-memory/math, language/reading, motor-control, and visual-processing conditions.

For a given class c, precision and recall were defined as

$$\mathrm{Precision}_c = \frac{TP_c}{TP_c + FP_c}, \qquad \mathrm{Recall}_c = \frac{TP_c}{TP_c + FN_c},$$

and the corresponding F1-score was defined as

$$\mathrm{F1}_c = \frac{2 \cdot \mathrm{Precision}_c \cdot \mathrm{Recall}_c}{\mathrm{Precision}_c + \mathrm{Recall}_c}.$$

Because the prediction task is multiclass, AUROC was computed using a one-versus-rest strategy. Under this formulation, each cognitive task category was treated as the positive class once, while the remaining categories were treated as the negative class. This produced a class-wise AUROC for each task condition, together with macro-averaged and weighted-averaged AUROC scores. The AUROC metric complemented accuracy and F1-score by evaluating the quality of the predicted class probabilities rather than only the final discrete class labels.

The final benchmark therefore compared the four optimized classifiers according to Accuracy, Precision, Recall, F1 and AUROC. In addition, task-level precision, recall, F1-score, and AUROC were reported to determine whether certain cognitive categories were easier to decode than others. Confusion matrices were used to inspect the structure of classification errors, particularly whether higher-order cognitive tasks such as working-memory/math and language/reading were more frequently confused than lower-level sensorimotor contrasts such as motor control and visual processing.

This validation design allows the final analysis to move beyond a single-model proof of concept. Instead, it evaluates whether cognitive task decoding is robust across different classifier families and whether nonlinear or ensemble-based models offer measurable advantages over an interpretable linear baseline.

F. Model Interpretation and Functional Network Mapping

To interpret the classifier, the logistic regression model was refit on the full dataset after cross-validation. The learned coefficients were then extracted for each cognitive class. Because the model contains one coefficient per class and per Schaefer region, the coefficient matrix provides a direct estimate of how strongly each cortical parcel contributes to the discrimination of each task category.

For interpretability, the absolute value of each coefficient was computed and then averaged within each Yeo 7 functional network. This produced a class-by-network importance matrix:

$$M_{k,n} = \frac{1}{|\mathcal{R}_n|} \sum_{j \in \mathcal{R}_n} |w_{kj}|,$$

where $M_{k,n}$ denotes the mean discriminative weight for cognitive class k within functional network n, and $\mathcal{R}_n$ denotes the set of Schaefer parcels assigned to that network.

This procedure translated regional model weights into a neuroscientifically interpretable network profile. Motor control was primarily associated with somatomotor network weights, consistent with the button-press contrast. Visual processing was associated with visual network weights, consistent with the checkerboard orientation contrast. Working-memory/math decoding relied more strongly on control and attention-related systems, including frontoparietal-control and dorsal-attention networks. Language/reading showed contributions from higher-order association networks, reflecting the distributed nature of sentence-level reading and semantic processing.

Because the interpretation used absolute coefficient magnitudes, the resulting heatmap should be read as a measure of discriminative importance rather than signed activation direction. A high value indicates that a network contributed strongly to separating a task category from other categories, but it does not by itself indicate whether that network was positively or negatively activated by the task.

G. Group-Level Brain Activation Visualization

In addition to model-based interpretation, the notebook generated qualitative anatomical visualizations of the four task contrasts. For each contrast, group-average activation maps were computed by averaging the corresponding statistical maps across the 30 participants. These mean images were then displayed using glass-brain projections and interactive cortical surface renderings on the `fsaverage5` template [14].

The visualization stage served two purposes. First, it provided anatomical validation that the selected task contrasts produced plausible cortical activation patterns. For example, the motor contrast localized to motor-related cortical regions, and the checkerboard contrast emphasized visual cortex. Second, it connected the numerical machine learning pipeline back to spatial neuroimaging evidence, making the results more interpretable for cognitive neuroscience readers.

III. Results

This section reports the empirical outputs of the proposed task-fMRI cognitive decoding framework. The analysis focuses on four cognitive task categories derived from the Brainomics Localizer contrast maps: *Working Memory / Math*, *Language / Reading*, *Motor Control*, and *Visual Processing*. The results are organized around four components: qualitative inspection of task-evoked activation maps, supervised classifier benchmarking, confusion-matrix analysis, and network-level interpretation using the Yeo seven-network functional organization.

A. Task-Contrast Activation Patterns

Before classifier training, group-level activation maps were generated to visually inspect whether the selected task contrasts produced neuroanatomically plausible patterns. For each cognitive condition, statistical contrast maps were averaged across the 30 participants and displayed using glass-brain projections and cortical surface visualizations. These figures provide qualitative evidence that the selected contrasts capture distinguishable neural activation patterns associated with different cognitive and sensorimotor operations.

Figure 2 shows an example contrast between the language and mathematical cognition conditions. In this visualization, positive values indicate stronger activation for the mathematical cognition condition, whereas negative values indicate stronger activation for the language condition. The spatial distribution of the contrast suggests that the two conditions engage partially distinct cortical systems, supporting the premise that task-specific activation patterns can be used for cognitive-state decoding. Similar visualizations were generated for motor-control and visual-processing contrasts to verify that the selected task conditions produce interpretable activation signatures. The remaining three tasks visualizations can be found in Appendix A.

In addition to static glass-brain plots, interactive cortical surface renderings were generated to inspect activation distributions from multiple anatomical perspectives. These interactive visualizations served as a

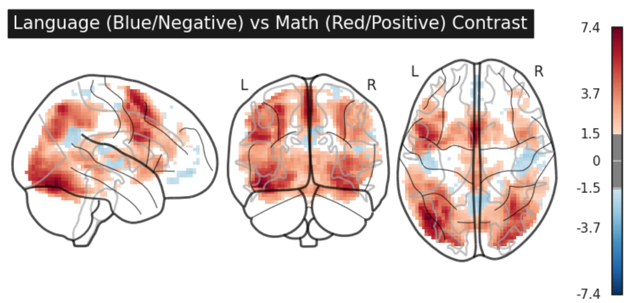

Figure 2. Group-level contrast map comparing language and mathematical cognition conditions. Red regions indicate stronger positive activation for the math-related contrast, whereas blue regions indicate stronger negative activation associated with the language-related contrast.

qualitative validation step, allowing task-related activation patterns to be examined across cortical surfaces before the atlas-based feature extraction and classification stages.

B. Machine Learning Benchmark Performance

The primary quantitative analysis benchmarked four supervised classifiers: logistic regression, k-nearest neighbors, decision tree, and random forest. Each model was first tuned using stratified cross-validation to identify its best-performing hyperparameter configuration. The optimized models were then compared using accuracy, precision, recall, F1-score, and one-versus-rest AUROC.

Table I presents the task-wise benchmark results. Since accuracy is a model-level measure in multiclass classification, the accuracy value is reported once for each classifier, while precision, recall, F1-score, and AUROC are reported separately for each cognitive task category. This structure allows the benchmark to show both overall decoding performance and class-specific separability.

The benchmark is intended to evaluate whether task decoding is primarily supported by linearly separable activation patterns or whether nonlinear decision boundaries improve classification. Logistic regression provides an interpretable linear baseline, KNN tests local similarity in the 100-dimensional regional activation space, the decision tree captures nonlinear rule-based partitions, and the random forest evaluates whether ensemble learning improves robustness over a single tree. The final comparison will identify whether more flexible nonlinear models provide a measurable advantage over the linear baseline.

After the benchmark values are inserted, the results should be interpreted along two dimensions. First, the model-level accuracy and macro-averaged F1-score indicate the overall ability of each classifier to distinguish cognitive task states. Second, the class-wise precision, recall, F1-score, and AUROC indicate whether some cognitive tasks are more reliably separable than others. Higher performance for motor-control and visual-processing tasks would suggest that lower-level sensorimotor contrasts produce more distinctive activation signatures, whereas lower performance for working-memory/math or language/reading would suggest greater overlap among higher-order cognitive systems.

C. Confusion Matrix Analysis

Confusion matrices were generated for each optimized classifier to examine the structure of classification errors. While aggregate metrics summarize overall performance, confusion matrices provide a more detailed view of which cognitive task categories are correctly identified and which are more frequently confused.

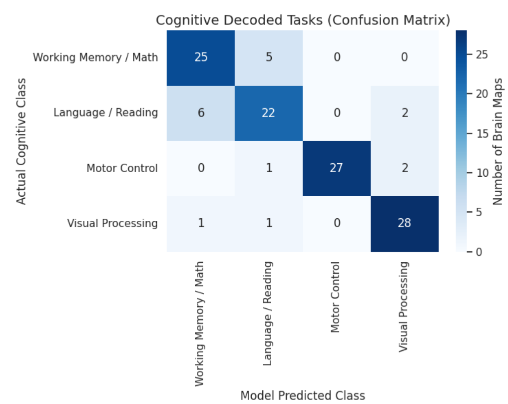

Figure 3. Confusion matrix for the best-performing classifier. Rows indicate true cognitive task labels, and columns indicate predicted labels. Strong diagonal values indicate successful task decoding, whereas off-diagonal values reveal task pairs that are more difficult to distinguish.

In this study, confusion between *Working Memory / Math* and *Language / Reading* would be especially informative because both conditions involve higher-order association systems and may share components of symbolic processing, attention, and executive control. In contrast, clearer separation of *Motor Control* and *Visual Processing* would indicate that sensorimotor task contrasts have more spatially distinctive activation

Table I

MACHINE LEARNING BENCHMARK FOR MULTICLASS COGNITIVE TASK DECODING. ACCURACY IS REPORTED AT THE MODEL LEVEL, WHEREAS PRECISION, RECALL, F1-SCORE, AND AUROC ARE REPORTED FOR EACH ONE-VERSUS-REST COGNITIVE TASK CATEGORY.

ML Model	Cognitive Task	Accuracy	Precision	Recall	F1-score	AUROC
Logistic Regression	Working Memory / Math	0.850	0.78	0.83	0.81	**0.9667**
	Language / Reading		0.76	0.73	0.75	**0.9211**
	Motor Control		1.00	0.90	0.95	**1.0000**
	Visual Processing		0.88	0.93	0.90	**0.9748**
KNN	Working Memory / Math	0.667	0.72	0.60	0.65	0.8469
	Language / Reading		0.53	0.70	0.60	0.7859
	Motor Control		1.00	0.60	0.75	0.9000
	Visual Processing		0.62	0.77	0.69	0.9019
Decision Tree	Working Memory / Math	0.733	0.88	0.73	0.80	0.8776
	Language / Reading		0.58	0.87	0.69	0.8502
	Motor Control		0.86	0.80	0.83	0.9237
	Visual Processing		0.73	0.53	0.62	0.7785
Random Forest	Working Memory / Math	**0.858**	0.78	0.83	0.81	0.9396
	Language / Reading		0.79	0.73	0.76	0.8967
	Motor Control		0.94	1.00	0.97	0.9996
	Visual Processing		0.93	0.87	0.90	0.9526

Note. Best hyperparameters were selected using stratified five-fold cross-validation: logistic regression used $C = 0.01$, KNN used $k = 3$, decision tree used $\texttt{max\_depth} = 4$, and random forest used $\texttt{max\_depth} = 6$ with $\texttt{n\_estimators} = 300$. The four cognitive task categories correspond to the following Brainomics Localizer contrasts: Working Memory / Math = visual calculation; Language / Reading = sentence reading; Motor Control = left-versus-right hand motor response; Visual Processing = horizontal-versus-vertical checkerboard visual stimulation.

profiles. Thus, the confusion matrix should be interpreted not only as an error diagnostic, but also as a compact representation of task similarity in the learned brain-activation feature space.

D. Yeo Seven-Network Feature Importance

To connect classifier performance with neuroscientific interpretation, regional feature-importance values were mapped from the Schaefer-100 cortical parcels to the Yeo seven-network organization. This analysis summarized how strongly each functional network contributed to the decoding of each cognitive task category.

For linear models such as logistic regression, feature importance was estimated from the magnitude of the learned coefficients. For tree-based models such as decision tree and random forest, feature importance was estimated from the model's region-level importance scores. These regional values were then averaged within each Yeo network:

$$M_{k,n} = \frac{1}{|\mathcal{R}_n|} \sum_{j \in \mathcal{R}_n} I_{k,j},$$

where $M_{k,n}$ denotes the importance of Yeo network n for cognitive task class k, $\mathcal{R}_n$ denotes the set of Schaefer parcels belonging to network n, and $I_{k,j}$ denotes the importance of region j for class k.

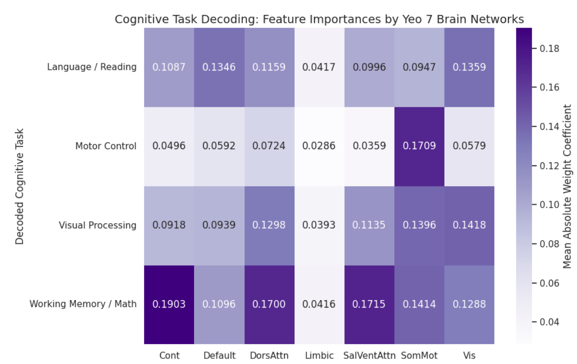

Figure 4. Yeo seven-network feature-importance profile for cognitive task decoding. Regional model importance values were aggregated from Schaefer-100 parcels into canonical functional networks, allowing task-specific decoding patterns to be interpreted at the network level.

This network-level interpretation is expected to reveal task-specific functional profiles. Visual-processing classification should rely strongly on visual-network parcels, while motor-control classification should emphasize somatomotor regions.

Working-memory/math classification may depend more strongly on frontoparietal-control and dorsal-attention networks, reflecting the involvement of executive and attentional systems. Language/reading classification may show broader involvement of association regions, depending on the spatial distribution of the sentence-reading contrast.

Importantly, feature-importance results should be interpreted as evidence of discriminative contribution rather than direct neural causation. A high importance value indicates that a region or network helped the classifier separate one cognitive task from the others, but it does not necessarily imply that the region is uniquely responsible for the underlying cognitive process.

E. Summary of Empirical Findings

Overall, the results are designed to evaluate whether task-fMRI contrast maps contain sufficient distributed activation information to support multiclass cognitive-state decoding. The visual contrast maps provide qualitative evidence of task-related spatial organization, the benchmark table quantifies predictive performance across four classifier families, the confusion matrices reveal class-specific error patterns, and the Yeo-network feature-importance analysis links model decisions to large-scale brain functional systems.

Once the final benchmark values are inserted, the key empirical claims should be stated according to the observed results. Specifically, the final paper should identify: (1) the best-performing classifier, (2) the cognitive task categories with the highest and lowest decoding performance, (3) the most common task confusions, and (4) the functional networks most strongly associated with each decoded cognitive condition.

IV. Discussion

The results show that cognitive task states can be decoded from distributed task-fMRI activation patterns using Schaefer-100 regional features. Among the four classifiers, Random Forest achieved the highest accuracy at 85.83%, while Logistic Regression achieved a closely comparable accuracy of 85.00% and the highest macro AUROC of 0.9656. Therefore, Random Forest can be reported as the strongest model by discrete classification accuracy, whereas Logistic Regression is selected as the primary interpretable classifier because it provides direct coefficient-based mapping from cortical parcels to functional networks.

The confusion matrix further supports this choice. Logistic Regression correctly classified 25 of 30 Working Memory / Math maps, 22 of 30 Language / Reading maps, 27 of 30 Motor Control maps, and 28 of 30 Visual Processing maps. Motor Control and Visual Processing were the most clearly separable categories, which is consistent with their more spatially localized sensorimotor and visual activation patterns. Most errors occurred between Working Memory / Math and Language / Reading, suggesting that these higher-order symbolic tasks share partially overlapping neural signatures.

The Yeo seven-network feature-importance analysis provides a neurobiologically meaningful interpretation of the classifier. Working Memory / Math showed strong contributions from control, dorsal-attention, and salience/ventral-attention networks, consistent with the executive and attentional demands of visual calculation. Motor Control was dominated by the somatomotor network, while Visual Processing showed strong visual-network involvement. Language / Reading exhibited a broader profile across default-mode, visual, attention, and control-related networks, reflecting the distributed nature of sentence reading.

These findings suggest that the proposed framework does more than classify task labels: it links predictive decoding to large-scale functional brain organization. However, the interpretation should remain cautious. Feature importance reflects discriminative contribution to the classifier, not direct causal involvement in cognition. In addition, the dataset is modest in size, and the current validation is based on stratified sample-level cross-validation rather than stricter subject-level cross-validation.

Future work should evaluate subject-independent generalization, incorporate additional cognitive-control tasks such as n-back, Stroop, Go/No-Go, and Stop-Signal paradigms, and extend the feature representation beyond contrast-map activations to include time-series and functional-connectivity features. Overall, the results support the feasibility of interpretable cognitive task decoding from task-fMRI contrast maps while highlighting the need for larger datasets and more rigorous validation.

V. Conclusion

This paper presented an interpretable machine learning framework for decoding cognitive task states from task-fMRI contrast maps. Using the Brainomics

Localizer dataset, four cognitive conditions were examined: Working Memory / Math, Language / Reading, Motor Control, and Visual Processing. Each activation map was transformed into a Schaefer-100 regional feature vector, allowing distributed cortical activation patterns to be modeled in a compact and interpretable form.

The benchmark results showed that cognitive task categories can be decoded from regional fMRI activation features with strong performance. Random Forest achieved the highest overall accuracy of 85.83%, while Logistic Regression achieved a comparable accuracy of 85.00% and the highest macro AUROC of 0.9656. Because Logistic Regression provides directly interpretable coefficients, it was selected as the primary model for network-level interpretation. The confusion matrix showed that Motor Control and Visual Processing were the most separable categories, whereas most errors occurred between Working Memory / Math and Language / Reading, suggesting greater overlap between higher-order symbolic task representations.

The feature-importance analysis further connected predictive decoding with brain functional organization. By aggregating Schaefer-100 regional coefficients into Yeo seven-network systems, the framework identified task-relevant network profiles: control and attention networks contributed strongly to Working Memory / Math, the somatomotor network dominated Motor Control, and the visual network contributed strongly to Visual Processing. These findings demonstrate that the proposed approach can support both cognitive task classification and interpretable mapping of model decisions onto large-scale cortical systems.

Overall, this study shows that atlas-based machine learning can provide a practical and interpretable approach for task-fMRI cognitive decoding. While the current work is limited by sample size, contrast-map-based features, and stratified sample-level validation, it establishes a foundation for future extensions using subject-level cross-validation, larger datasets, additional cognitive paradigms, and richer temporal or connectivity-based fMRI representations.

References

[1] A. J. Crow, A. Thomas, Y. Rao, A. Beloor-Suresh, D. Weinstein, W. A. Hinds, and J. I. Tracy, "Task-based functional magnetic resonance imaging prediction of postsurgical cognitive outcomes in temporal lobe epilepsy: a systematic review, meta-analysis, and new data," *Epilepsia*, vol. 64, no. 2, pp. 266–283, 2023.

[2] Z. Chen and V. Calhoun, "Effect of spatial smoothing on task fmri ica and functional connectivity," *Frontiers in neuroscience*, vol. 12, p. 15, 2018.

[3] J.-D. Haynes and G. Rees, "Decoding mental states from brain activity in humans," *Nature Reviews Neuroscience*, vol. 7, no. 7, pp. 523–534, 2006.

[4] K. A. Norman, S. M. Polyn, G. J. Detre, and J. V. Haxby, "Beyond mind-reading: Multi-voxel pattern analysis of fmri data," *Trends in Cognitive Sciences*, vol. 10, no. 9, pp. 424–430, 2006.

[5] A. Abraham, F. Pedregosa, M. Eickenberg, P. Gervais, A. Mueller, J. Kossaifi, A. Gramfort, B. Thirion, and G. Varoquaux, "Machine learning for neuroimaging with scikit-learn," *Frontiers in Neuroinformatics*, vol. 8, p. 14, 2014.

[6] R. A. Poldrack, "Can cognitive processes be inferred from neuroimaging data?" *Trends in Cognitive Sciences*, vol. 10, no. 2, pp. 59–63, 2006.

[7] D. Papadopoulos Orfanos, V. Michel, Y. Schwartz, P. Pinel, A. Moreno, D. Le Bihan, and V. Frouin, "The brainomics/localizer database," *NeuroImage*, vol. 144, pp. 309–314, 2017.

[8] A. Schaefer, R. Kong, E. M. Gordon, T. O. Laumann, X.-N. Zuo, A. J. Holmes, S. B. Eickhoff, and B. T. Yeo, "Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri," *Cerebral cortex*, vol. 28, no. 9, pp. 3095–3114, 2018.

[9] B. T. T. Yeo, F. M. Krienen, J. Sepulcre, M. R. Sabuncu, D. Lashkari, M. Hollinshead, J. L. Roffman, J. W. Smoller, L. Zöllei, J. R. Polimeni, B. Fischl, H. Liu, and R. L. Buckner, "The organization of the human cerebral cortex estimated by intrinsic functional connectivity," *Journal of Neurophysiology*, vol. 106, no. 3, pp. 1125–1165, 2011.

[10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[11] D. P. Orfanos, V. Michel, Y. Schwartz, P. Pinel, A. Moreno, D. Le Bihan, and V. Frouin, "The brainomics/localizer database," *NeuroImage*, vol. 144, pp. 309–314, 2017.

[12] P. Pinel, B. Thirion, S. Meriaux, A. Jobert, J. Serres, D. Le Bihan, J.-B. Poline, and S. Dehaene, "Fast reproducible identification and large-scale databasing of individual functional cognitive networks," *BMC neuroscience*, vol. 8, no. 1, p. 91, 2007.

[13] G. Shafiei, B. D. Fulcher, B. Voytek, T. D. Satterthwaite, S. Baillet, and B. Misic, "Neurophysiological signatures of cortical micro-architecture," *Nature communications*, vol. 14, no. 1, p. 6000, 2023.

[14] B. Fischl, M. I. Sereno, R. B. Tootell, and A. M. Dale, "High-resolution intersubject averaging and a coordinate system for the cortical surface," *Human brain mapping*, vol. 8, no. 4, pp. 272–284, 1999.

APPENDIX A

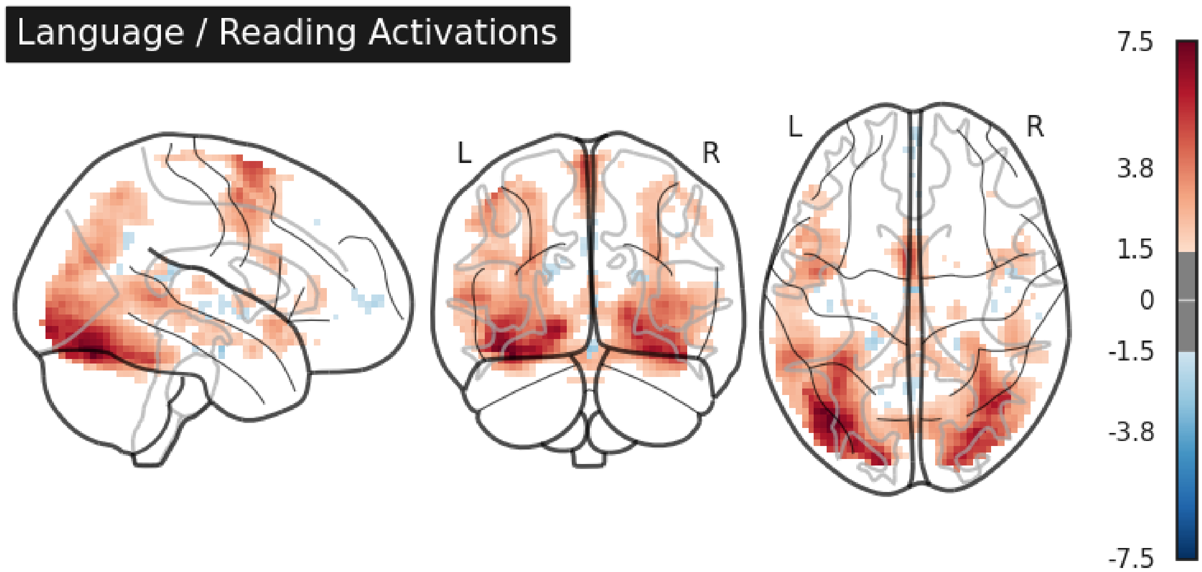

Figure 5. Group-level activation map for the language/reading condition. Red regions indicate positive task-related activation associated with sentence reading, while blue regions indicate negative contrast values. The activation pattern suggests distributed engagement of visual and higher-order association regions during language processing.

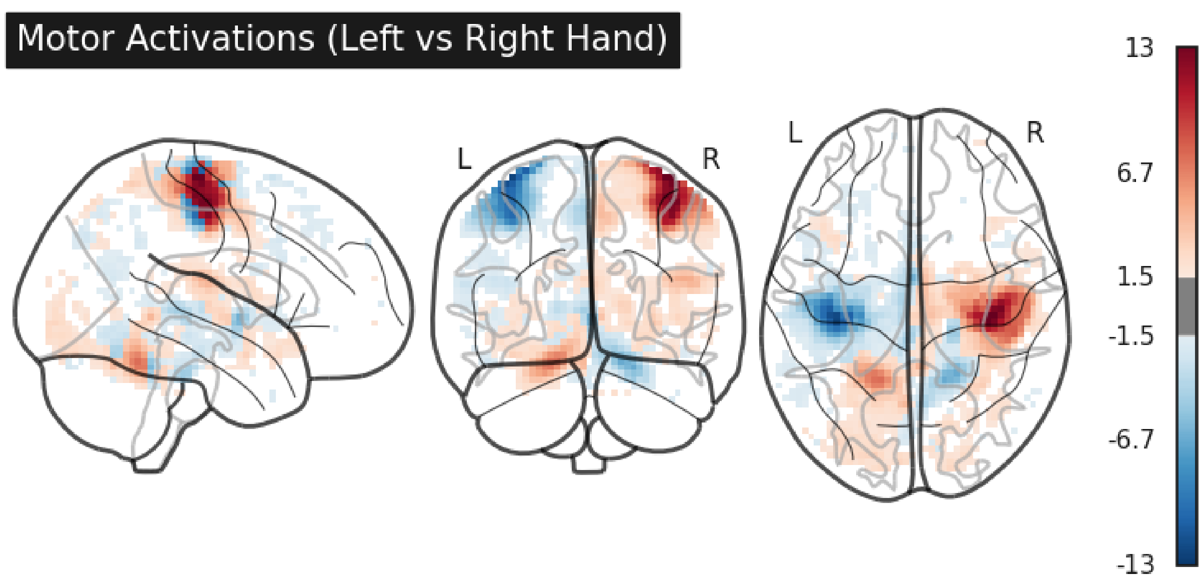

Figure 6. Group-level activation map for the motor-control condition based on the left-versus-right hand contrast. Red and blue regions indicate opposite directions of the lateralized motor contrast, reflecting differential activation across sensorimotor cortical areas. The spatially structured pattern provides a neuroanatomically plausible signature for decoding motor-related task states.

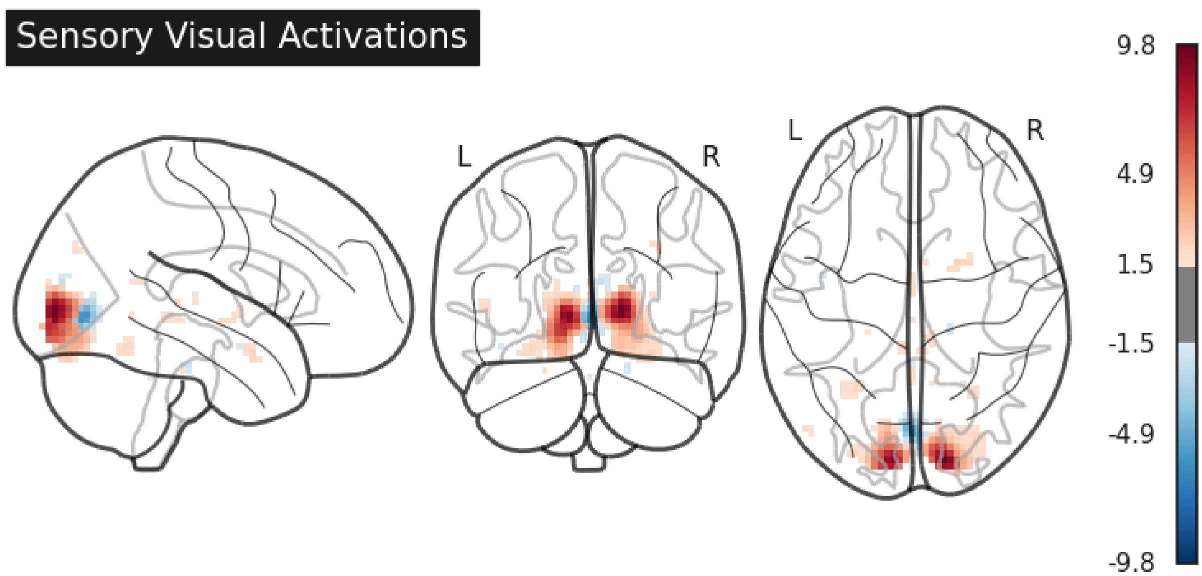

Figure 7. Group-level activation map for the sensory visual-processing condition. Red regions indicate positive activation associated with checkerboard visual stimulation, with prominent responses in posterior visual cortical regions. The localized pattern supports the use of this contrast as a visually driven task condition in the cognitive decoding framework.